Rational optimization for nonlinear reconstruction with approximate $\ell 0$ penalization

Marc Castella, Jean-Christophe Pesquet, Arthur Marmin

Rational Optimization for Nonlinear Reconstruction with Approximate ℓ_0 Penalization

Marc Castella, *Member, IEEE*, Jean-Christophe Pesquet, *Fellow, IEEE*, and Arthur Marmin

***Abstract*—Recovering nonlinearly degraded signal in the presence of noise is a challenging problem. In this work, this problem is tackled by minimizing the sum of a non convex least-squares fit criterion and a penalty term. We assume that the nonlinearity of the model can be accounted for by a rational function. In addition, we suppose that the signal to be sought is sparse and a rational approximation of the ℓ_0 pseudo-norm thus constitutes a suitable penalization. The resulting composite cost function belongs to the broad class of semi-algebraic functions. To find a globally optimal solution to such an optimization problem, it can be transformed into a generalized moment problem, for which a hierarchy of semidefinite programming relaxations can be built. Global optimality comes at the expense of an increased dimension and, to overcome computational limitations concerning the number of involved variables, the structure of the problem has to be carefully addressed. A situation of practical interest is when the nonlinear model consists of a convolutive transform followed by a componentwise nonlinear rational saturation. We then propose to use a sparse relaxation able to deal with up to several hundreds of optimized variables. In contrast with the naive approach consisting of linearizing the model, our experiments show that the proposed approach offers good performance.**

***Index Terms*—signal reconstruction, sparse signal, nonlinear model, polynomial optimization, semi-definite programming.**

I. Introduction

OVER the last decade, there has been much progress made in the area of sparse signal recovery. The results and techniques have spread over a wide range of signal processing applications such as denoising, source separation, image restoration, or image reconstruction. Attention has been however mostly focused on linear observation models, for which many efforts have been dedicated to solving the associated inverse problems. In the basic setup, a vector of observations $\mathbf{d}$ is available, which is obtained from a ground-truth signal $\overline{\mathbf{x}}$ by a linear transformation $\mathbf{H}$. It is known that an exact reconstruction of $\overline{\mathbf{x}}$ is possible even when the size of the latter is greater than the number of observations, a fact popularized by the celebrated compressed sensing theory [1] which exploits the structure (i.e. sparsity) of $\overline{\mathbf{x}}$.

Unfortunately, the linear assumption on the observation model is often quite inaccurate. For a long time and in many signal processing applications, attempts have been made in order to deal with more general nonlinear models. For example, one can mention the pioneering works undertaken with Volterra models [2], which may be useful in some application areas [3]. More recently, the work in [4] has explicitly taken into account a nonlinearity, but the reconstruction results hold under restrictive assumptions. Similarly, for many real acquisition devices, the actual degradation model is not linear as some nonlinear saturation effects often arise. This situation is closely related to 1-bit compressed sensing [5] and classification problems. Such nonlinearly distorted convolution models may also be encountered in blind source separation [6] and neural networks [7]. A simplified model resulting from a linearization procedure can then be adopted in order to make the associated mathematical problem tractable. For example, standard tools in signal processing such as the Wiener filter are effective mostly in a linear framework. More specifically, well-known sparse recovery techniques such as LASSO have been used in a nonlinear context by overlooking the nonlinearity. Some results have been obtained in this context [8], [9], [10], but methods explicitly taking into account the nonlinearity are likely to provide better results and are crucially lacking. This paper aims at providing such a method in this still unexplored area.

A popular approach in many reconstruction problems consists in minimizing the sum of a data fidelity term and a regularization term incorporating prior information such as sparsity. In this case, convex potentials related to the ℓ_1 norm are often employed as surrogates to the natural sparsity measure, which is the ℓ_0 pseudo-norm (count of the number of nonzero components in the signal). Although some theoretical works have promoted the use of the ℓ_1 norm [1], its optimality can only been established under some restrictive assumptions. In turn, cost functions involving the ℓ_0 pseudo-norm lead to NP-hard problems for which reaching a global minimum cannot be guaranteed in general [11], [12], [13]. Smooth approximations of the ℓ_0 pseudo-norm may appear to provide good alternative solutions [14], [15], [16], [17]. Among the class of possible smoothed ℓ_0 functions, the Geman-McClure $\ell_2-\ell_0$ potential was observed to give good results in a number of applications [14], [15], [16]. Yet, in the recent works [18], [19], [20], [17], promising results have been obtained with a non differentiable function.

Concerning the minimization of the penalized criterion, many efforts have been undertaken to derive efficient algorithms able to deal with a large number of variables, while ensuring convergence to a global minimizer [21], [22], [23]. Many of the available techniques assume that the observation model is linear and that the noise has a log-concave likelihood. Then, both the penalty and the data fit terms are convex and

M. Castella (corresponding author) is with SAMOVAR, Télécom SudParis, CNRS, Université Paris-Saclay, 9 rue Charles Fourier, 91011 Evry Cedex, France. E-mail: marc.castella@telecom-sudparis.eu.

J.-C. Pesquet and A. Marmin are with the Center for Visual Computing, CentraleSupélec, INRIA and Université Paris-Saclay, 91192 Gif sur Yvette, France. E-mail: jean-christophe@pesquet.eu, arthur.marmin@centralesupelec.fr.

many optimization techniques may be used. In a more difficult scenario, a quadratic tangent function can be derived, which makes efficient majorization-minimization (MM) strategies usable for optimizing certain penalized criteria (see [24] for more details). However, for most of the existing optimization algorithms (e.g. those based on Majorize-Minimize strategies), only convergence to a local minimum can be expected and algorithms can get trapped by undesirable local minima due to the nonconvexity of the criterion. In our context, none of the two terms of the criterion is convex because of the nonlinear observation model and because of the chosen approximation of the ℓ_0 pseudo-norm. Developing methods with global convergence properties is therefore a crucial issue, which we address in this paper.

An approach recently proposed in the optimization community [25], [26], [27], [28] provides theoretical guarantees of global optimality when only polynomial or rational functions are involved. The minimization problem is recast as a problem of moments, for which a hierarchy of semidefinite positive programming (SDP) relaxations asymptotically provides an exact solution. This approach is often referred to as the Lasserre hierarchy [25] and its major advantage is a guaranteed convergence to the global minimum of the original problem, which can be accessed by solving successive SDP problems. Alternatively, the problem of global polynomial or rational minimization can be tackled from the standpoint of sum of squares (SOS) hierarchy [26], [28], [29]: both approaches are linked by duality. One advantage of the moment approach is the possibility, under some conditions, to extract the optimal point.

We investigate here the potential offered by these rational optimization methods for sparse signal recovery from nonlinearly degraded observations. In the present state of research, the Lasserre/SOS hierarchies are restricted to small to medium size problems. In signal processing, one of the main difficulties we face is the large number of variables which have to be optimized. A stochastic block-coordinate method has been proposed as a first solution in one of our previous works [18]: despite interesting experimental results, global optimality is lost in this case.

In this work, we propose a novel approach for restoring sparse signals degraded by a nonlinear model. More precisely, our contributions in this paper are threefold.

- First, the proposed approach is able to deal with degradation models consisting of a convolution followed by a pointwise transform. The latter appears as a rational fraction of the absolute value of its input argument. The formulation of the problem as a nonconvex optimization also allows the use of a Geman-McClure like regularization term.
- Although SDP relaxations of optimization problems are popular in signal processing [30], they usually lead to suboptimal solutions. Our second contribution is to make use of asymptotically exact SDP relaxations able to minimize polynomial or rational nonconvex functions of several variables.
- The last contribution of this work is to devise a sparse relaxation in the spirit of [31] to cope with the resulting rational optimization. Exploiting the specific structure of the problem to obtain sparse SDP relaxations plays a prominent role in making the Lasserre/SOS hierarchy applicable to several hundred of variables as it is common in inverse problems.

The remainder of the paper is organized as follows. The considered model is described in Section II. Section III describes the general methodology and Section IV emphasizes the specificities of our context. Simulations results are provided in Section V. Finally, Section VI concludes the paper.

Notation: The set of polynomials in the indeterminates given by vector $\mathbf{x} = (x_1, \ldots, x_T) \in \mathbb{R}^T$ is denoted by $\mathbb{R}[\mathbf{x}]$. For any multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T) \in \mathbb{N}^T$, we define $\mathbf{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \ldots x_T^{\alpha_T}$ and $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_n$. Therefore, any polynomial can be written as a finite sum over multi-indices as follows: $(\forall \mathbf{x} \in \mathbb{R}^T)\ p(\mathbf{x}) = \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}$. The degree of p will be denoted by $\deg p$. Such a polynomial can be identified with the vector of its coefficients $\mathbf{p} = (p_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}}$: this will be used for convenience. Finally, the lowest integer upper bound of any real-valued number is denoted by $\lceil . \rceil$.

II. Model and criterion

A. Observation model

We consider the problem of recovering a set of unknown samples given by the vector $\overline{\mathbf{x}} := (\overline{x}_1, \ldots, \overline{x}_T)^\top$. In our context, this original signal cannot be measured and we have access only to some measurements related to the original signal through a linear transformation followed by some nonlinear effects. More precisely, the observation model reads

$$\mathbf{d} = \boldsymbol{\phi}(\mathbf{H}\overline{\mathbf{x}}) + \mathbf{n}, \tag{1}$$

where the vector $\mathbf{d} = (d_1, \ldots, d_T)^\top$ contains the observation samples, $\mathbf{n} = (n_1, \ldots, n_T)^\top$ is a perturbation noise vector, $\mathbf{H} \in \mathbb{R}^{T \times T}$ is a given matrix, and $\boldsymbol{\phi} : \mathbb{R}^T \to \mathbb{R}^T$ is a nonlinear function. It is assumed that $\boldsymbol{\phi}$ applies componentwise, that is, for every $\mathbf{u} := (u_1, \ldots, u_T)^\top$, the t-th component of $\boldsymbol{\phi}(\mathbf{u})$ is given by $[\boldsymbol{\phi}(\mathbf{u})]_t = \phi_t(u_t)$, where the real-valued function ϕ_t models a saturation effect as in the top plot of Figure 1. In this paper, the functions $(\phi_t)_{1 \leq t \leq T}$ are assumed to be known and to be rational, possibly involving absolute values. Actually, $(\phi_t)_{1 \leq t \leq T}$ being saturation is only a practical example and other functions could be considered in theory as long as the function $\boldsymbol{\phi}$ in (1) is semi-algebraic (see the comments in Subsection III-B6).

The linear part in Model (1) can typically describe a convolution. When the matrix $\mathbf{H}$ is Toeplitz band as in Section IV-A, with values defined by the finite impulse response $(h_t)_t$ of a filter, and with vanishing boundary conditions, the samples in (1) indeed stem from a signal given by

$$(\forall t \in \{1, \ldots, T\}) \quad d_t = \phi_t(h_t \star \overline{x}_t) + n_t. \tag{2}$$

In the equation above, $\star$ denotes the sequence convolution and $\big(n_t\big)_{1 \leq t \leq T}$ is a realization of an additive random noise. An important contribution in this paper is that this structure will be exploited in order to reduce the computational cost of the subsequently proposed global optimization method (see Section IV). For now, no assumption is made on the matrix $\mathbf{H}$.

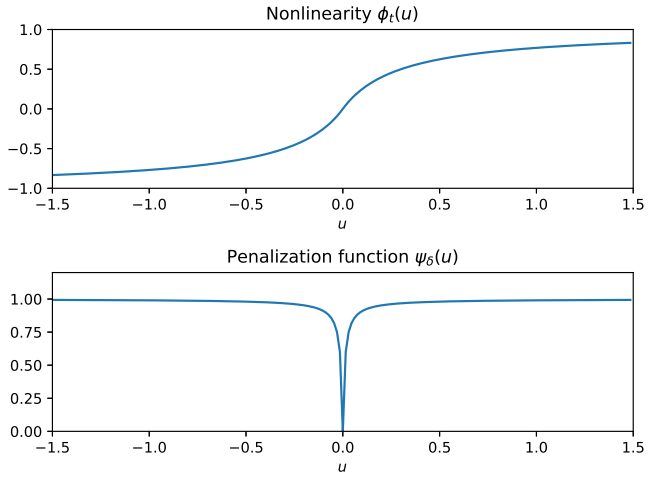

Fig. 1: Plot of the nonlinear saturation function ϕ_t in (28) with $\chi = 0.3$ and of the sparsity promoting function in (4) with $\delta = 0.01$.

B. Sparse signal and penalized criterion

The signal $(\overline{x}_t)_{1\leq t\leq T}$ modelled by vector $\overline{\mathbf{x}}$ is assumed to be sparse. By saying this, we simply assume that $\overline{x}_t \neq 0$ only for a few indices t.

Following a classical approach for estimating $\overline{\mathbf{x}}$, we minimize a penalized criterion having the following form:

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad \mathcal{J}(\mathbf{x}) = \|\mathbf{d} - \boldsymbol{\phi}(\mathbf{H}\mathbf{x})\|^2 + \lambda \mathcal{P}(\mathbf{x})\,, \tag{3}$$

where $\mathcal{P}$ is a penalization function whose small values promote sparse vectors, in accordance with our assumptions concerning the true $\overline{\mathbf{x}}$. The positive regularization parameter λ controls the relative importance given to the squared norm fit term and to the penalization. In this paper, we have chosen $\mathcal{P}(\mathbf{x}) = \sum_{t=1}^{T} \psi_\delta(x_t)$ where the sparsity promoting function ψ_δ has been drawn on the bottom plot of Figure 1 and is given by

$$(\forall \xi \in \mathbb{R}) \qquad \psi_\delta(\xi) = \frac{|\xi|}{\delta + |\xi|}\,. \tag{4}$$

This choice is similar in spirit to the Geman-McClure potential [14] and, since for every $\xi \in \mathbb{R}$, $\lim_{\delta\to 0} \psi_\delta(\xi) = 0$ when $\xi = 0$ and 1 otherwise, the solution to the ℓ_0 penalized problem is recovered asymptotically as $\delta \to 0$ under some technical assumptions (see [15, Proposition 2]). Note also that this penalty has recently shown to be effective in image restoration problems (see [32] and references therin). Finally, the criterion to be minimized in our approach reads:

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad \mathcal{J}(\mathbf{x}) = \|\mathbf{d} - \boldsymbol{\phi}(\mathbf{H}\mathbf{x})\|^2 + \lambda \sum_{t=1}^{T} \psi_\delta(x_t)\,. \tag{5}$$

The minimization is performed over a feasible set $\mathbf{K}$, which is assumed to be compact. This assumption is required later in (11) and (14) and it makes no restriction since the signal values $\overline{\mathbf{x}}$ are bounded in practice. The optimal cost function value is denoted by

$$\mathcal{J}^\star = \inf_{\mathbf{x}\in\mathbf{K}} \mathcal{J}(\mathbf{x})\,,$$

and the estimated signal generated by our approach is then $\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbf{K}} \mathcal{J}(\mathbf{x})$.

Importantly, our model involves rational functions $\boldsymbol{\phi}$ and ψ_δ. As a consequence, the criterion $\mathcal{J}$ is a rational function of its arguments (possibly with absolute values). We detail in next section how recent rational and polynomial optimization techniques apply in this context.

III. Polynomial and rational optimization

In this section, we explain the basic principles of the methods in [25], [27] for polynomial and rational optimization.

A. Global optimization and equivalent problem over measure

Consider the problem of determining the *global* minimum of a given lower-semicontinuous function f over a given compact set $\mathbf{K} \subset \mathbb{R}^T$:

$$\text{Find} \quad f^\star = \inf_{\mathbf{x}\in\mathbf{K}} f(\mathbf{x}). \tag{6}$$

We can introduce an optimization problem equivalent to (6), where the new optimization variable is a measure belonging to an infinite dimensional space. Following the terminology from [27], such problem will be called a *generalized problem of moments* (GPM). Denoting by $\mathcal{P}(\mathbf{K})$ the set of probability measures suppported on $\mathbf{K}$, this problem reads as follows:

$$\text{Find} \quad (f^\star)_{\text{gpm}} = \inf_{\mu\in\mathcal{P}(\mathbf{K})} \int_{\mathbb{R}^T} f(\mathbf{x})\; d\mu(\mathbf{x})\,. \tag{7}$$

To see the equivalence between (6) and (7), note first that $(\forall \mathbf{x} \in \mathbf{K})\ f(\mathbf{x}) \geq f^\star$ implies that $(f^\star)_{\text{gpm}} \geq f^\star$. For the reverse inequality, it can be noticed that the minimum of f is reached at a point $\mathbf{x}^\star \in \mathbf{K}$ because $\mathbf{K}$ is compact and the Dirac measure $\delta_{\mathbf{x}^\star}$ provides a solution such that $(f^\star)_{\text{gpm}} = f^\star$.

Let us now write more specifically the GPM for rational functions by assuming that the function f reads:

$$(\forall \mathbf{x} \in \mathbb{R}^T)\ \ f(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})} \quad \text{where} \quad (\forall \mathbf{x} \in \mathbf{K})\ \ q(\mathbf{x}) > 0\,, \tag{8}$$

where p and q are polynomials. Let us introduce the measure $d\nu(\mathbf{x}) = \frac{1}{q(\mathbf{x})} d\mu(\mathbf{x})$. With this change of variables, ν is no longer a probability measure, but, since the total mass of the probability measure μ is one, it satisfies (10) below. Therefore, by defining $\mathcal{M}(\mathbf{K})$ as the set of finite nonnegative Borel measures supported on $\mathbf{K}$, Problem (7) can be equivalently re-expressed as:

$$\text{Find} \quad f^\star = \inf_{\nu\in\mathcal{M}(\mathbf{K})} \int_{\mathbb{R}^T} p(\mathbf{x})\; d\nu(\mathbf{x}) \tag{9}$$

$$\text{s.t.} \int_{\mathbb{R}^T} q(\mathbf{x}) d\nu(\mathbf{x}) = 1\,. \tag{10}$$

Importantly, Problem (9)-(10) corresponds to a simple objective function and an explicit constraint, both terms being linear with respect to ν. However, the implicit contraint that ν is a measure in $\mathcal{M}(\mathbf{K})$ is more complicated to cope with. Fortunately, the latter can be handled via a hierarchy of tractable constraints when p and q are polynomials, as shown next.

B. Hierarchy of SDP relaxations

The infinite dimensional optimization problem (9) – (10) can be approximated by a hierarchy of SDP problems with increasing sizes when the involved function is given by (8) with $(p, q) \in (\mathbb{R}[\mathbf{x}])^2$. The main ingredients of this approach are presented now.

1) Moment sequence: In (9)-(10), the optimization variable is the measure ν. The first step is to replace this variable by a more tractable one, i.e. a finite dimensional vector. Since the measure ν has a compact support, it can be represented by a moment sequence $\mathbf{y} = (y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha} \in \mathbb{N}^T}$ defined as

$$(\forall \boldsymbol{\alpha} \in \mathbb{N}^T) \quad y_{\boldsymbol{\alpha}} = \int_{\mathbf{K}} \mathbf{x}^{\boldsymbol{\alpha}} \, d\nu(\mathbf{x}). \tag{11}$$

In so doing, the measure ν in Problem (9)-(10) is represented by the moment sequence $\mathbf{y}$, which is an infinite dimensional vector. A hierarchy of finite dimensional optimization problems will be obtained by considering truncated versions of $\mathbf{y}$ with increasing sizes.

2) Linear objective and constraints: Consider a polynomial of total degree k represented by its vector of coefficients $\mathbf{p} = (p_{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}| \leq k}$:

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad p(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| \leq k} p_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}}. \tag{12}$$

By linearity and by the definition of the moments $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}}$, any integral such as the ones arising in (9) and (10) can be rewritten as

$$\int_{\mathbb{R}^T} p(\mathbf{x}) \, d\nu(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| \leq k} p_{\boldsymbol{\alpha}} y_{\boldsymbol{\alpha}} = \mathcal{L}_p(\mathbf{y}). \tag{13}$$

The function $\mathcal{L}_p(\cdot)$ as defined above is linear. Therefore, the objective and the explicit constraint in (9) and (10) are linear funtions of the moment sequence $\mathbf{y}$ and the difficulty of the original problem has therefore been transfered to the implicit constraint that the sequence $\mathbf{y}$ should satisfy (11) for a given measure $\nu \in \mathcal{M}(\mathbf{K})$.

3) Support/measure constraint: Since an arbitrary sequence $\mathbf{y}$ does not necessarily represent a measure ν on $\mathbf{K}$, some constraints needs to be taken into account on $\mathbf{y}$. To achieve this goal, we first need a more precise description of $\mathbf{K}$. In our context, the set $\mathbf{K}$ is defined by polynomial inequalities of the following form:

$$\mathbf{K} = \{\mathbf{x} \in \mathbb{R}^T \mid (\forall i \in \{1, \ldots, I\}) \, g_i(\mathbf{x}) \geq 0\}, \tag{14}$$

where, for every $i \in \{1, \ldots, I\}$, $g_i \in \mathbb{R}[\mathbf{x}]$. The constraints will now be specified on a truncated version of the sequence $\mathbf{y}$. For a given $k \in \mathbb{N}$ and for multi-indices $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that $|\boldsymbol{\alpha}| \leq k$ and $|\boldsymbol{\beta}| \leq k$, the elements of the k-th order *moment matrix* $\mathbf{M}_k(\mathbf{y})$ of $\mathbf{y}$ are given by

$$[\mathbf{M}_k(\mathbf{y})]_{\boldsymbol{\alpha}, \boldsymbol{\beta}} = y_{\boldsymbol{\alpha} + \boldsymbol{\beta}}. \tag{15}$$

Note that $\mathbf{M}_k(\mathbf{y})$ involves moments up to the order $2k$. The main property of $\mathbf{M}_k(\mathbf{y})$ is that for a polynomial of degree no more than k expressed by (12), we have:

$$\int_{\mathbf{K}} p(\mathbf{x})^2 \, d\nu(\mathbf{x}) = \mathbf{p}^\top \mathbf{M}_k(\mathbf{y}) \mathbf{p}. \tag{16}$$

Similarly, for a given polynomial $g \in \mathbb{R}[\mathbf{x}]$, the elements of the *localizing matrix* $\mathbf{M}_k^g(\mathbf{y})$ associated to g and $\mathbf{y}$ are

$$[\mathbf{M}_k^g(\mathbf{y})]_{\boldsymbol{\alpha}, \boldsymbol{\beta}} = \sum_{\boldsymbol{\gamma}} g_{\boldsymbol{\gamma}} y_{\boldsymbol{\gamma} + \boldsymbol{\alpha} + \boldsymbol{\beta}}, \tag{17}$$

and we have

$$\int_{\mathbf{K}} g(\mathbf{x}) p(\mathbf{x})^2 \, d\nu(\mathbf{x}) = \mathbf{p}^\top \mathbf{M}_k^g(\mathbf{y}) \mathbf{p}. \tag{18}$$

The positivity of the right hand side of (16) for any vector of coefficients $\mathbf{p}$ shows that the positive semi-definiteness of matrix $\mathbf{M}_k(\mathbf{y})$ is a necessary condition for the sequence $\mathbf{y}$ to be a valid moment sequence. Similarly, according to (18) and because $(\forall \mathbf{x} \in \mathbf{K}) \; g_i(\mathbf{x}) \geq 0$, we deduce that $\mathbf{M}_k^{g_i}(\mathbf{y}) \succeq 0$ for every $i \in \{1, \ldots, I\}$, if $\mathbf{y}$ is the moment sequence of a measure in $\mathcal{M}(\mathbf{K})$. Due to the linear dependence of $\mathbf{M}_k(\mathbf{y})$ and $\mathbf{M}_k^g(\mathbf{y})$ on $\mathbf{y}$, these constraints are linear matrix inequalities.

4) Relaxation: Based on the above developments, we are now able to introduce a relaxation of Problem (9)-(10). Define, for every $i \in \{1, \ldots, I\}$, $r_i = \lceil (\deg g_i)/2 \rceil$ and, for any order $k \geq \max\{\max_{i=1}^I r_i, \deg p, \deg q\}$, consider the optimization problem:

$$\begin{aligned} \text{Find} \quad f_k^\star = \inf_{\mathbf{y}} \; & \mathcal{L}_p(\mathbf{y}) \\ \text{s.t.} \; & \mathcal{L}_q(\mathbf{y}) = 1, \\ & \mathbf{M}_k(\mathbf{y}) \succeq 0, \\ & \mathbf{M}_{k - r_i}^{g_i}(\mathbf{y}) \succeq 0 \quad (\forall i \in \{1, \ldots, I\}). \end{aligned} \tag{19}$$

The objective function and the equality constraint are directly derived from Problem (9)-(10) where the integrals have been represented as in (13). The last two constraints are necessary constraints for $\mathbf{y}$ to be a measure supported by $\mathbf{K}$. Therefore, we naturally have $f_k^\star \leq f^\star$ and $f_k^\star$ is an increasing sequence with lower bounds of $f^\star$. Note that since the order in the last constraints have been limited to $k - r_i$, for every $i \in \{1, \ldots, I\}$, it follows from (15) and (17) that the moments involved in Problem (19) are $(y_{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}| \leq 2k}$.

A crucial observation is that (19) is a convex SDP optimization problem for which efficient techniques exist and provide guaranteed global optimal solution [33], [34].

5) Theoretical results and solution extraction:

a) Convergence results: We now detail some existing theoretical results about the approach. For their validity, the following technical assumption is required:

A1. There exist polynomials $\sigma_0, \ldots, \sigma_I$, which are all sum of squares, such that the set $\{\mathbf{x} \in \mathbb{R}^T \,|\, \sigma_0(\mathbf{x}) + \sum_{i=1}^I \sigma_i(\mathbf{x}) g_i(\mathbf{x}) \geq 0\}$ is compact.

Under Assumption A1, we have [25], [27]

$$f_k^\star \uparrow f^\star \text{ as } k \to +\infty \, .$$

This is a strong result ensuring convergence to the *global* optimum of Problem (6) when considering increasing order SDP relaxations.

Note that, in addition to $\mathbf{K}$ being compact, Condition A1 requires that that the polynomials $(g_i)_{1 \leq i \leq I}$ describing $\mathbf{K}$ in (14) yield an algebraic certificate of compactness. More details can be found in [26], [27], [28], [31]. For simplicity, we will

only consider the practical situation where $\mathbf{K} = [\underline{B}, \overline{B}]^T$. This is easily satisfied when lower and upper bounds $\underline{B}, \overline{B}$ on the variables $(\overline{x_t})_{1\leq t\leq T}$ are available. By setting $I = T$ and

$$(\forall \mathbf{x} \in \mathbb{R}^T) \qquad g_t(\mathbf{x}) = (x_t - \underline{B})(\overline{B} - x_t),$$

$\mathbf{K}$ can be expressed under the form (14). The set $\mathbf{K}$ is obviously compact. In addition,

$$\begin{aligned} \sum_{t=1}^{T} g_t(x_t) \geq 0 \;\; &\Leftrightarrow \;\; \|\mathbf{x}\|^2 - (\overline{B} + \underline{B}) \sum_{t=1}^{T} x_t + T\underline{B}\overline{B} \leq 0 \\ &\Leftrightarrow \;\; \|\mathbf{x} - \mathbf{u}\| \leq T\frac{\overline{B} - \underline{B}}{2}, \end{aligned} \tag{20}$$

where $\mathbf{u} = \frac{\overline{B}+\underline{B}}{2}(1, \ldots, 1)^\top \in \mathbb{R}^T$. Therefore, Assumption A1 holds with $\sigma_0(\mathbf{x}) = 0$ and, for every $t \in \{1, \ldots, T\}$ $\sigma_t(\mathbf{x}) = 1$.

b) Extraction of the optimal solution: The above convergence results are asymptotic results for increasing orders of the hierarchy of SDP relaxations. Fortunately, it has been experimentally observed that low relaxation orders often provide satisfactory results (see e.g. [25], [31]). In addition, it has been proven [27], [26] that under certain rank conditions, the solution given by the SDP relaxation is guaranteed to be the global minimizer of the original problem. In this case, globally optimal points can be extracted by the procedure in [35]. Details on the rank conditions and the extraction procedure are beyond the scope of this paper.

Unfortunately, there are two main difficulties in applying this methodology to practical situations: first, it is known that detecting the rank of a matrix can be numerically sensitive. In addition, because of the complexity of the original problem, the possible relaxation order that we can choose may be too small. For both reasons, we have observed that the mentioned rank conditions are generally not satisfied numerically. Alternatively, considering that the global minimium is likely to be unique, one can extract from the optimal solution $\mathbf{y}^\star$ to Problem (19) the moments corresponding to the respective monomials $x_1, \ldots, x_T$. This extraction is straightforward and we have used the vector of these moments as an estimate denoted by $\hat{\mathbf{x}}_k^\star$ of the global minimizer for the original problem.

6) Extension to semi-algebraic functions/constraints: From a theoretical viewpoint, the above methodology can be extended to more complicated functions and constraints than polynomials or fractions. We briefly explain how the absolute value, which appears in the nonlinearity and/or in the penalty function, can be handled. Proceeding similarly, it is actually possible to handle any semi-algebraic function or constraint.

First, note that polynomial equality constraints such as $g(\mathbf{x}) = 0$ are possible in the definition of the feasible set $\mathbf{K}$. This is easily done by introducing the two inequalities $g(\mathbf{x}) \geq 0$ and $-g(\mathbf{x}) \geq 0$ in the equations defining $\mathbf{K}$ in (14).

Then, absolute values can be considered as follows: for each term $|v(\mathbf{x})|$ appearing, where v is a polynomial, one can introduce an additional variable u and impose the constraints $u \geq 0, u^2 = v(\mathbf{x})^2$. The methodology of the paper can then be applied with the extended set of variables $(\mathbf{x}, u)$.

IV. Exploiting the problem structure

A. Toeplitz structure and split criterion

In this section, we assume that the convolutional model in (2) is considered. Additionally, it is assumed that the involved filter is FIR with impulse response of length L given by the vector $(h_1, \ldots, h_L)^\top$. Under vanishing boundary conditions, the observation model in (1) holds and involves the following specific Toeplitz band matrix:

$$\mathbf{H} = \begin{bmatrix} h_1 & 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ h_L & & & \ddots & & \vdots \\ 0 & \ddots & & & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \ldots & 0 & h_L & \ldots & h_1 \end{bmatrix}.$$

Finally, remind that we have assumed that the nonlinearity ϕ applies componentwise and that it is given by a rational function, possibly involving absolute values. The latters can be discarded by using the trick described above. Thus, for clarity, and with no loss of generality, we describe the method when all quantities are nonnegative and hence no absolute value appears.

We now focus on two specificities of our problem and show how they can leverage a methodology similar to [31]. Note that the methodology remains applicable when a subsampling is performed on the observation vector $\mathbf{y}$ (see [36]). First, developing the squared norm in (5) and substituting all terms, the criterion $\mathcal{J}$ appears as a sum of rational functions. Reducing $\mathcal{J}$ to the same denominator would result in a ratio of high degree polynomials, making the approach described in Section III intractable. A remedy consists in introducing one measure (and hence one moment sequence) for each elementary fraction in $\mathcal{J}$, and simultaneously imposing constraints which guarantee equality of identical moments related to different measures.

Going further, the second specificity stems from the Toeplitz band structure of the matrix $\mathbf{H}$. In this case indeed, each term of the sum of rational functions in $\mathcal{J}(\mathbf{x})$ only involves a small subset of all variables. This leads to a sparse[1] SDP relaxation. The rationale is explained below and more details are given in Section IV-B.

Let us introduce, for every $t \in \{1, \ldots, T\}$, the set

$$I_t = \{\min\{1, t - L + 1\}, \ldots, t\}$$

which is the set of column indices where t-th row of $\mathbf{H}$ has nonzero elements (in particular, $I_1 = \{1\}$, $I_2 = \{1, 2\}$, ..., $I_T = \{T - L + 1, \ldots, T\}$). Developing the squared norm, we rewrite Criterion (5) as follows

$$\mathcal{J}(\mathbf{x}) = \sum_{t=1}^{T} \underbrace{\left(d_t - \phi_t\Big(\sum_{i=1}^{L} h_i x_{t-i+1} \Big) \right)^2}_{\frac{p_{I_t}(\mathbf{x})}{q_{I_t}(\mathbf{x})}} + \underbrace{\lambda \psi_\delta(x_t)}_{\frac{p(x_t)}{q(x_t)}}, \tag{21}$$

[1]The notion of sparsity here concerns the optimization variables and should not be confused with the sparsity assumed for the original samples in vector $\mathbf{x}$.

where by convention $x_t = 0$ for every $t \notin \{1,\dots,T\}$. This reads equivalently:

$$\mathcal{J}(\mathbf{x}) = \sum_{t=1}^{T} \left(\frac{p_{I_t}(\mathbf{x})}{q_{I_t}(\mathbf{x})} + \frac{p(x_t)}{q(x_t)} \right). \tag{22}$$

In the above equation, p_{I_t}, q_{I_t} are polynomials that depend on the variables $(x_k)_{k\in I_t}$ only and $p(x_t), q(x_t)$ are univariate polynomials that depend on x_t only.

Now, one can see that, by introducing for each fraction summing up in (21) a relaxation similar to the methodology introduced in Section III, the original problem involving a large number T of variables is split in a collection of smaller problems and relaxations. Proceeding in this way would be quite natural for a separable criterion where the problem is decomposed into a sum of subproblems that can be solved independently. Of course, for a non separable criterion, one cannot split freely the problem and constraints must be added between the subproblems to link them. In addition, a technical condition is required on the subsets of variables of the split form. This is further explained in the next section.

B. Sparse SDP relaxation

For every $t \in \{1,\dots,T\}$, each rational function $\frac{p_{I_t}(\mathbf{x})}{q_{I_t}(\mathbf{x})}$ is related to the marginal μ_{I_t} on $\mathbb{R}^{|I_t|}$ of the original probability measure μ defined on $\mathbb{R}^T$. By weighting μ_{I_t} with the denominator of this rational fraction, as explained in Section III, we define a measure ν_{I_t} associated with a sequence of moments $\mathbf{z}_t$, which satisfies the following relations: for any $k \geq \max\{1, \deg p_{I_t}, \deg q_{I_t}\}$,

$$\mathbf{M}_k(\mathbf{z}_t) \succeq 0, \;\; \mathcal{L}_{q_{I_t}}(\mathbf{z}_t) = 1, \;\; \mathbf{M}^{g_t}_{k-r_t}(\mathbf{z}_t) \succeq 0. \tag{23}$$

In addition, we have to pay attention to the fact that the same monomial may appear in consecutive terms $\frac{p_{I_{t-1}}(\mathbf{x})}{q_{I_{t-1}}(\mathbf{x})}$ and $\frac{p_{I_t}(\mathbf{x})}{q_{I_t}(\mathbf{x})}$ in Summation (22), when $t \in \{2,\dots,T\}$. Let $\mathbb{N}^{(I_t\cap I_{t-1})}$ denote the subset of T-tuples $\boldsymbol{\alpha} = (\alpha_1,\dots,\alpha_T) \in \mathbb{N}^T$ such that $\alpha_t = 0$ for $t \notin I_t \cap I_{t-1}$. In other words, the T-tuples in $\mathbb{N}^{(I_t\cap I_{t-1})}$ correspond to monomials involving variables with indices in $I_t \cap I_{t-1}$. The latter monomials are precisely the common monomials in $\frac{p_{I_{t-1}}(\mathbf{x})}{q_{I_{t-1}}(\mathbf{x})}$ and $\frac{p_{I_t}(\mathbf{x})}{q_{I_t}(\mathbf{x})}$. We have then, for every $\boldsymbol{\alpha} \in \mathbb{N}^{(I_t\cap I_{t-1})}$,

$$\begin{aligned} &\int \mathbf{x}^{\boldsymbol{\alpha}} d\mu_{I_t}(\mathbf{x}) = \int \mathbf{x}^{\boldsymbol{\alpha}} d\mu_{I_{t-1}}(\mathbf{x}) \\ \Leftrightarrow \quad &\mathcal{L}_{\mathbf{x}^{\boldsymbol{\alpha}} q_{I_t}(\mathbf{x})}(\mathbf{z}_t) = \mathcal{L}_{\mathbf{x}^{\boldsymbol{\alpha}} q_{I_{t-1}}(\mathbf{x})}(\mathbf{z}_{t-1}). \end{aligned} \tag{24}$$

Similarly, for every $t \in \{1,\dots,T\}$, the rational function $\frac{p(x_t)}{q(x_t)}$ can be associated with a sequence of monovariate moments $\mathbf{y}_t$, for which the following conditions have to be met:

$$\mathbf{M}_k(\mathbf{y}_t) \succeq 0, \;\; \mathcal{L}_{q_t}(\mathbf{y}_t) = 1, \;\; \mathbf{M}^{g_t}_{k-r_t}(\mathbf{y}_t) \succeq 0, \tag{25}$$

and, for every $\alpha \in \mathbb{N}$,

$$\mathcal{L}_{x_t^\alpha q(x_t)}(\mathbf{y}_t) = \mathcal{L}_{x_t^\alpha q_{I_t}(\mathbf{x})}(\mathbf{z}_t). \tag{26}$$

By using these variables $(\mathbf{y}_t, \mathbf{z}_t)_{1\leq t\leq T}$, we are now in order to provide a sparse SDP relaxation for the minimization of (22):

$$\begin{aligned} \text{Find } f_k^{\star \mathsf{s}} = &\inf_{\mathbf{z},\mathbf{y}} \sum_{t=1}^{T} \mathcal{L}_{p_{I_t}}(\mathbf{z}_t) + \mathcal{L}_p(\mathbf{y}_t) \\ &\text{s.t. } (\forall t \in \{1,\dots,T\}): \\ &\quad (23), (25), \\ &\quad (24) \text{ for } \boldsymbol{\alpha} \in \mathbb{N}^{(I_t\cap I_{t-1})} \text{ with } |\boldsymbol{\alpha}| + \deg q_{I_t} \leq 2k, \\ &\quad (26) \text{ for } \alpha + \deg q_{I_t} \leq 2k. \end{aligned}$$

Remark 1: For the aforementioned approach to be mathematically valid, a technical assumption is required: the so-called Running Intersection Property [31], [27]. For convenience, let us introduce a notation for the $2T$ different index sets corresponding to each fraction in (22):

$$(\forall t \in \{1,\dots,T\}) \quad J_t = I_t \text{ and } J_{t+T} = \{t\}.$$

Note that the sets $(J_t)_{1\leq t\leq 2T}$ satisfy $\bigcup_{t=1}^{2T} J_t = \{1,\dots,T\}$. The Running Intersection Property then reads

$$(\forall t \in \{2,\dots,2T\})\; J_t \bigcap \left(\bigcup_{k=1}^{t-1} J_k \right) \subseteq J_j \quad \text{for some } j \leq t-1. \tag{27}$$

It is easy to check that this condition is satisfied in our case.

C. Comparison between full and sparse relaxations

We detail here the reasons why the specific form of the latter relaxation is crucial from a computational standpoint. Using the sparse relaxation indeed allows us to handle a much higher number of variables T than the non sparse one. The different numbers of involved variables and matrix sizes are listed below, in the case when no absolute value appears.

1) Relaxation involving one measure only: For a problem with T variables and a relaxation order k, the size of the vector representing the measure/moment sequence is given by the number of all mutivariate monomials with degree less than or equal to $2k$, which is precisely the binomial coefficient $\binom{T+2k}{2k}$. As a consequence, the number of variables in an SDP relaxation involving only one measure (such as (19)) scales as T^{2k}. In addition, according to the definition of the moment matrix in (15), the maximum size of the square matrices defining positive definite constraints is $\binom{T+k}{k}$, which scales as T^k.

2) Sparse relaxation for a Toeplitz matrix: Concerning the sparse relaxation with order k, the number of variables involved is $\binom{L+2k}{2k}$ for each $\mathbf{z}_t$ and $\binom{1+2k}{2k} = 2k+1$ for each $\mathbf{y}_t$ with $t \in \{1,\dots,T\}$. The total number of variables in the sparse SDP relaxation is therefore

$$T\left(\binom{L+2k}{2k} + 2k + 1 \right).$$

As a consequence, for a given order k, the number of variables scales as $T\,L^{2k}$ in the computation of $f_k^{\star \mathsf{s}}$. The maximum size

of the moment matrix with positive definite constraint is then $\binom{L+k}{k}$, hence it scales as L^k.

In summary, the gain in terms of size of the sparse relaxation is T^{2k-1}/L^{2k}. In addition, the maximum size of the semidefinite constraints is of the respective order T^k for the non sparse relaxation and L^k for the sparse one. Considering these two facts, it follows that the sparse relaxation is highly advantageous for $L \ll T$, that is for $\mathbf{H}$ corresponding to a convolutive matrix with relatively short FIR.

Remark 2: The relaxation order k must be greater than or equal to the maximal degree appearing in the original polynomial or rational problem. Consequently, Relaxation (19) is intractable after reducing the terms in (22) to the same denominator, since this would introduce polynomials with a high degree (of order T). On the contrary, the sparse relaxation takes explicitly into account that the criterion is a sum of fractions with low degrees and allows order k to be set to a much smaller value.

V. Simulations

A. Experimental setup

1) Generated sparse signal and nonlinearity: In all the performed experiments, several sets of 100 Monte-Carlo realizations of generated data are processed. Samples $\overline{\mathbf{x}}$ of a sparse signal are generated, the number of samples being set to $T = 200$, $T = 100$, $T = 50$, or $T = 20$. We impose that exactly 10% of the sample values are nonzero, yielding respectively 20, 10, 5, and 2 nonzero components in $\overline{\mathbf{x}}$.

Then, this impulsive signal, considered as the ground truth, is corrupted following the model in (1), where the noise $\mathbf{n}$ is drawn according to an i.i.d. zero-mean Gaussian distribution with standard deviation $\sigma = 0.15$. The components of the nonlinear function $\boldsymbol{\phi}$ are chosen all identical and given by

$$(\forall t \in \{1, \ldots, T\}) \quad \phi_t(u) = \frac{u}{\chi + |u|}, \tag{28}$$

where $\chi = 0.3$. Considering the amplitude of the signals, the above function acts as a nonlinear saturation (see top plot in Figure 1). Finally, the matrix $\mathbf{H}$ is Toeplitz band and corresponds to FIR filters of length 3.

We test our approach in two scenarios:

a) Nonnegative case: We first consider only a nonnegative original signal $\overline{\mathbf{x}}$ and nonnegative coefficients in the matrix $\mathbf{H}$. In the Geman-McClure penalty term given by (4), absolute value are then of no use and they can be discarded. The amplitudes of the nonzero components of $\overline{\mathbf{x}}$ are drawn according to a uniform distribution on $[2/3, 1]$. The impulse responses of the FIR filters corresponding to $\mathbf{H}$ are set to $\mathbf{h}^{(a)} = [0.1, 0.8, 0.1]$ or $\mathbf{h}^{(b)} = [0.2254, 0.3361, 0.4385]$. An additional set of Monte-Carlo simulations is run where the impulse responses are drawn randomly (nonnegatively) for each realization. Due to the positivity assumption, the minimization of $\mathcal{J}^\star$ is then performed on the hypercube $\mathbf{K} = [0, 1]^T$.

b) Real-valued case: We then consider real valued $\overline{\mathbf{x}}$ and $\mathbf{H}$, still using the penalty term in (4). The amplitudes of the nonzero components of $\overline{\mathbf{x}}$ are then drawn according to a uniform distribution on $[-1, -2/3] \cup [2/3, 1]$. In addition, the impulse responses of the FIR filters are given by $\mathbf{h}^{(a)}$, $\mathbf{h}^{(b)}$ (like in the first scenario), and $\mathbf{h}^{(c)} = [-0.1127, -0.0683, 0.8191]$. Here again, on one set of Monte-Carlo realizations, the impulse responses are randomly drawn with real-valued coeffcients, for each realization. Finally, the criterion is minimized on the set $\mathbf{K} = [-1, 1]^T$.

2) Considered optimization methods: Recall that the optimized criterion is given by (5). In both scenarios, we have set empirically $\lambda = 0.15$ for the regularization parameter and $\delta = 0.01$ in the penalty function (4).

To obtain an estimate of $\overline{\mathbf{x}}$, we have built the sparse SDP relaxation from Section IV-B with orders $k = 2$ and $k = 3$ using the software [37]. The SDP has then been solved using SDPT3 [33]. Finally, the corresponding estimate $\mathbf{x}_k^{\star \mathbf{s}}$ is determined as described in Section III-B.

We are not aware of any other method able to find the global minimum of (5). For comparison with a globally convergent approach, we have used a linearized model for reconstruction purposes: based on Model (1), we have linearized around zero the nonlinearity (28) and have used the well-known ℓ_1 penalization. The cost function then reads

$$(\forall \mathbf{x} \in \mathbb{R}^T) \quad \mathcal{J}_{\ell_1}(\mathbf{x}) = \left\| \mathbf{d} - \frac{1}{\chi}\mathbf{H}\mathbf{x} \right\|^2 + \lambda_1 \sum_{t=1}^{T} |x_t|, \ \lambda_1 > 0$$

and it can thus be minimized efficiently by standard convex optimization techniques [21], [38].

Finally, we have also implemented a proximal gradient algorithm corresponding to the well-known Iterative Hard Thresholding (IHT) [12]. Since the function $\boldsymbol{\phi}$ is Lipschitz-differentiable, the standard IHT algorithm can be extended to the nonlinear observation model. This leads to the following iterative algorithm:

$$(\forall n \in \mathbb{N}) \ \mathbf{x}^{(n+1)} = \mathbf{S}_{\sqrt{\lambda_0 \eta}}\Big(\mathbf{x}^{(n)} - \eta \mathbf{H}^\top \nabla \boldsymbol{\phi}(\mathbf{H}\mathbf{x}^{(n)}) \big(\boldsymbol{\phi}(\mathbf{H}\mathbf{x}^{(n)}) - \mathbf{d}\big)\Big)$$

where the Jacobian matrix $\nabla \boldsymbol{\phi}(\mathbf{H}\mathbf{x}^{(n)})$ is diagonal and $\mathbf{S}_{\sqrt{\lambda_0 \eta}}$ is the hard thresholder with threshold value $\sqrt{\lambda_0 \eta}$, $\lambda_0 > 0$. It can be shown that any value of the stepsize η in $]0, \eta_{\max}]$ is valid, where $1/\eta_{\max} = \|\mathbf{H}\|_{\mathrm{S}}^2 (1 + 2 \max_{1 \le t \le T} |d_t|)/\chi^2$ is a Lipschitz constant of the above gradient term ($\|\mathbf{H}\|_{\mathrm{S}}$ denotes the spectral norm of $\mathbf{H}$). The latter algorithm however only certifies convergence to a local minimum of the criterion [39]. Due to non convexity, the local minima are likely to differ from the global minimum [11].

B. Results

1) Performance of the proposed relaxation : Figures 2 and 3 show the objective values $\mathcal{J}(\mathbf{x}_k^{\star \mathbf{s}})$ and the lower-bounds $f_k^{\star \mathbf{s}}$ provided by our method for relaxation orders $k = 2$ and $k = 3$, and for two different sample sizes. The value of the objective function $\mathcal{J}$ obtained after minimizing $\mathcal{J}_{\ell_1}$ is also plotted. For readability, the Monte-Carlo realizations have been sorted by

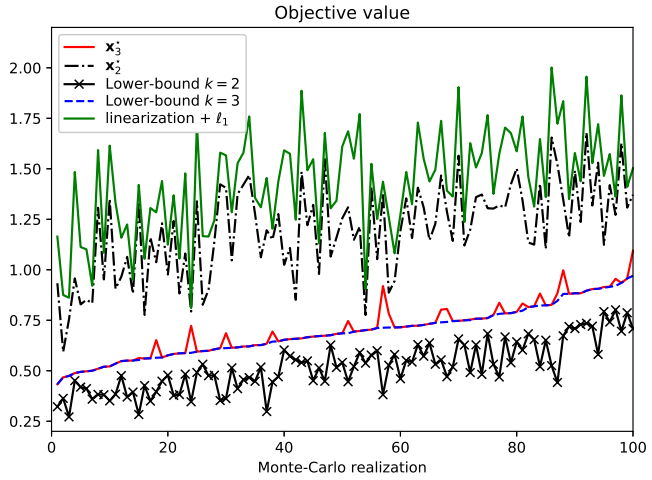

Fig. 2: Objective value and lower-bound given by our method (randomly driven filters, nonnegative case, $T = 20$).

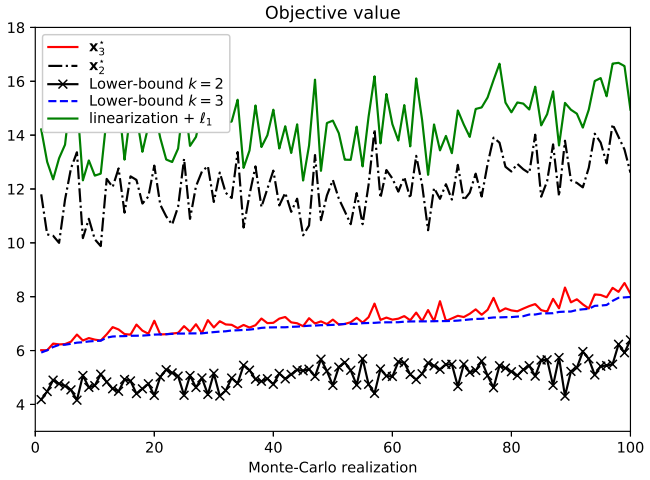

Fig. 3: Objective value and lower-bound given by our method (randomly driven filters, nonnegative case, $T = 200$).

increasing value of $f_3^{\star \mathbf{s}}$. The poor performance of the convex formulation may be accounted for by the fact that the linearized model leads to a rough approximation. In accordance with the theory, we have $f_2^{\star \mathbf{s}} \leq f_3^{\star \mathbf{s}}$ and the latter value is indeed a lower-bound on the corresponding obtained criterion values, which are obviously such that $\mathcal{J}(\mathbf{x}_3^{\star \mathbf{s}}) \leq \mathcal{J}(\mathbf{x}_2^{\star \mathbf{s}})$. Moreover, the gap between $f_k^{\star \mathbf{s}}$ and $\mathcal{J}(\mathbf{x}_k^{\star \mathbf{s}})$ is an evidence of the effectiveness of our method. A strictly positive value, as observed for $k = 2$ indicates that the relaxation order is too small. As illustrated in Figures 2 and 3 the gap value reduces for $k = 3$, and with $T = 20$ a gap numerically close to zero certifies that the global minimum is perfectly attained in more than 80% of the cases. For the more involved case $T = 200$, the gap value is small with $k = 3$ but nonzero: this gives evidence in favor of closeness to the global solution, although a higher relaxation order would probably be necessary. Due to memory limitations, increasing further the relaxation order is unfortunately impossible so far. In the next section, we show how to combine our method with IHT, so as to alleviate this issue.

2) Dealing with local minimas: Because of the difficulty of the rational optimization task, we propose to complement our method with the IHT optimization method, which is known to be efficient, but only locally. For better emphasizing the benefit of our approach, several initializations of IHT are considered: $\mathbf{x}_3^{\star \mathbf{s}}$, the result from the linearized model and ℓ_1 penalization, $\mathbf{d}$, an all-zero vector, and the true $\overline{\mathbf{x}}$. Obviously, the latter initialization would be impossible to use in real applications. The average values over all Monte-Carlo realizations are provided in Tables I (nonnegative case) and II (real-valued case) for $T = 200$. Some more detailed plots, corresponding to randomly drawn filter coefficients, are shown in Figures 4 (nonnegative case) and 5 (real-valued case).

The final objective values after convergence of IHT clearly depend on the initialization, which witnesses the existence of several local optima and emphasizes the importance of addressing the problem from a global optimization standpoint. In average, the lowest objective value is obtained by a local

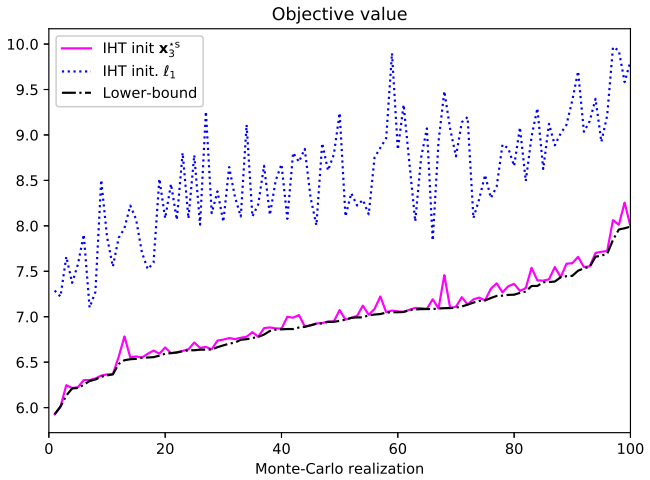

Fig. 4: Objective value for IHT with different initializations (randomly driven filters, nonnegative case, $T = 200$). .

optimization initialized either at $\mathbf{x}_3^{\star \mathbf{s}}$ or at the true $\overline{\mathbf{x}}$, the two choices leading to very similar results. More importantly, as shown in Tables III (nonnegative case) and IV (real-valued case), IHT is not reliable for finding the global minimum. These two tables compare different initializations of IHT and provide for each initialization the number of times it leads to the smallest objective value among the 100 Monte-Carlo realizations (a sum greater than 100 on a row occurs for $T = 20$ and indicates that different initializations have reached the same minimum value). In the overwhelming majority of cases, the initialization with $\mathbf{x}_3^{\star \mathbf{s}}$ provides the smallest objective. As soon as T is more than a few tens, IHT is almost unable to reach the global minimum with any of the standard initializations (ℓ_1, $\mathbf{d}$, all-zero vector). This demonstrates the fact that the proposed relaxation is useful in providing a good initial point for a local optimization algorithm.

3) Signal recovery performance: Finally, we illustrate the merits of our method in terms of estimation and peak detection

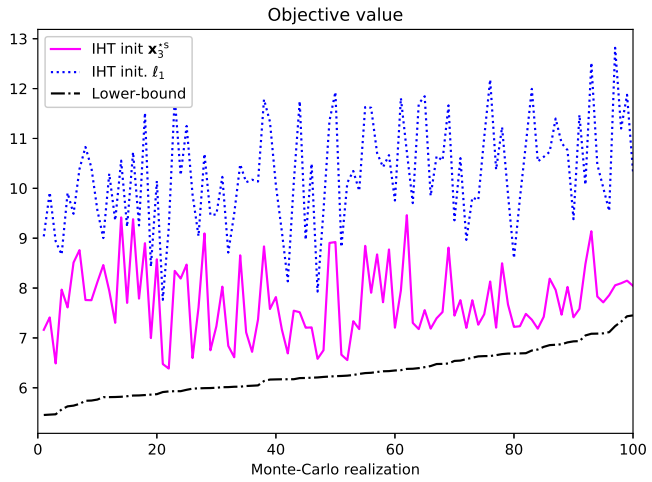

Fig. 5: Objective value for IHT with different initializations (randomly driven filters, real-valued case, $T = 200$). .

TABLE I: Final values of the objective function $\mathcal{J}$ for various optimization methods (nonnegative case, $T = 200$).

Opt. method	Filters		
	$\mathbf{h}^{(a)}$	$\mathbf{h}^{(b)}$	random
$\mathbf{x}_3^{\star s}$	7.3185	7.1317	7.1528
linearized ℓ_1	15.749	13.794	14.406
IHT, init. $\mathbf{x}_3^{\star s}$	7.0970	7.0424	6.9981
IHT, init. ℓ_1	8.7043	8.6388	8.5518
IHT, init. $\mathbf{d}$	8.8508	8.8928	9.1245
IHT, init. zero	11.798	10.014	13.988
IHT, init. $\overline{\mathbf{x}}$	7.1441	7.1476	7.1060

TABLE II: Final values of the objective function $\mathcal{J}$ for various optimization methods (real-valued case, $T = 200$).

Opt. method	Filters			
	$\mathbf{h}^{(a)}$	$\mathbf{h}^{(b)}$	$\mathbf{h}^{(c)}$	random
$\mathbf{x}_3^{\star s}$	12.0845	17.3860	12.2985	16.389
linearized ℓ_1	21.837	20.0003	21.7529	20.786
IHT, init. $\mathbf{x}_3^{\star s}$	7.2254	8.2095	7.2131	7.7278
IHT, init. ℓ_1	10.048	11.7268	9.3964	10.281
IHT, init. $\mathbf{d}$	10.024	11.2028	11.9485	12.934
IHT, init. zero	12.079	15.5946	10.4484	12.8234
IHT, init. $\overline{\mathbf{x}}$	7.1323	7.1113	7.1363	7.1151

errors. A typical example of true signal $\overline{\mathbf{x}}$, of observation vector $\mathbf{d}$ and of reconstructed signal is displayed in Figure 6. The estimation error on $\overline{\mathbf{x}}$ has been quantified by the mean square error $\frac{1}{T}\|\hat{\mathbf{x}} - \overline{\mathbf{x}}\|^2$ for a given estimate $\hat{\mathbf{x}}$. The average error and objective values are gathered in Tables V (nonnegative case) and VI (real-valued case). It can be observed that the results obtained with the ℓ_1 penalization followed by IHT are significantly improved when the initialization of IHT is performed by the proposed rational optimization approach.

Finally, we have compared our method for detecting the peaks in the original signal. Nonzero values of $\overline{\mathbf{x}}$ have been estimated by comparing $|\hat{\mathbf{x}}|$ to a threshold. The so-called receiver operating characteristic (ROC) curves are plotted on Figure 7 by increasing the threshold value: it represents the detection rate versus the false alarm rate. Clearly, using $\mathbf{x}_3^{\star s}$ gives the best results. On the contrary, the linearized model

TABLE III: Out of 100 Monte-Carlo realizations, number of times each initialization of IHT provides the smallest objective value (nonnegative case, filter random (top) and $\mathbf{h}^{(a)}$ (bottom)).

Num. samples	Initialization			
	$\mathbf{x}_3^{\star s}$	ℓ_1	$\mathbf{d}$	zero
random filter				
20	87	6	4	11
50	100	0	0	0
100	100	0	0	0
200	100	0	0	0
filter $\mathbf{h}^{(a)}$				
20	86	1	4	17
50	99	0	0	1
100	100	0	0	0
200	100	0	0	0
filter $\mathbf{h}^{(b)}$				
20	94	6	4	5
50	100	0	0	0
100	100	0	0	0
200	100	0	0	0

TABLE IV: Out of 100 Monte-Carlo realizations, number of times each initialization of IHT provides the smallest objective value (real-valued case, filters $\mathbf{h}^{(a)}$ and $\mathbf{h}^{(b)}$).

Num. samples	Initialization			
	$\mathbf{x}_3^{\star s}$	ℓ_1	$\mathbf{d}$	zero
random filter				
20	74	7	6	18
50	97	0	1	2
100	99	1	0	0
200	100	0	0	0
filter $\mathbf{h}^{(a)}$				
20	79	2	5	18
50	100	0	0	0
100	100	0	0	0
200	100	0	0	0
filter $\mathbf{h}^{(b)}$				
20	87	2	7	4
50	100	0	0	0
100	100	0	0	0
200	100	0	0	0
filter $\mathbf{h}^{(c)}$				
20	62	6	8	32
50	97	1	0	2
100	99	0	0	1
200	100	0	0	0

TABLE V: Final average MSE for the proposed optimization method (nonnegative case, $T = 200$).

Opt. method	Filters		
	$\mathbf{h}^{(a)}$	$\mathbf{h}^{(b)}$	random
IHT, init. $\mathbf{x}_3^{\star s}$	9.23e-03	1.16e-2	1.12e-2
IHT, init. ℓ_1	1.17e-02	1.42e-2	1.34e-2
IHT, init. $\mathbf{d}$	1.73e-02	1.43e-2	1.59e-2
IHT, init. zero	5.06e-02	6.47e-2	5.89e-2

TABLE VI: Final average MSE for the proposed optimization method (real-valued case, $T = 200$).

Opt. method	Filters			
	$\mathbf{h}^{(a)}$	$\mathbf{h}^{(b)}$	$\mathbf{h}^{(c)}$	random
IHT, init. $\mathbf{x}_3^{\star s}$	9.50e-3	1.58e-2	9.27e-3	1.08e-2
IHT, init. ℓ_1	1.35e-2	3.09e-2	1.22e-2	1.73e-2
IHT, init. $\mathbf{d}$	2.66e-2	2.91e-2	4.43e-2	3.34e-2
IHT, init. zero	5.30e-2	6.66e-2	4.23e-2	5.17e-2

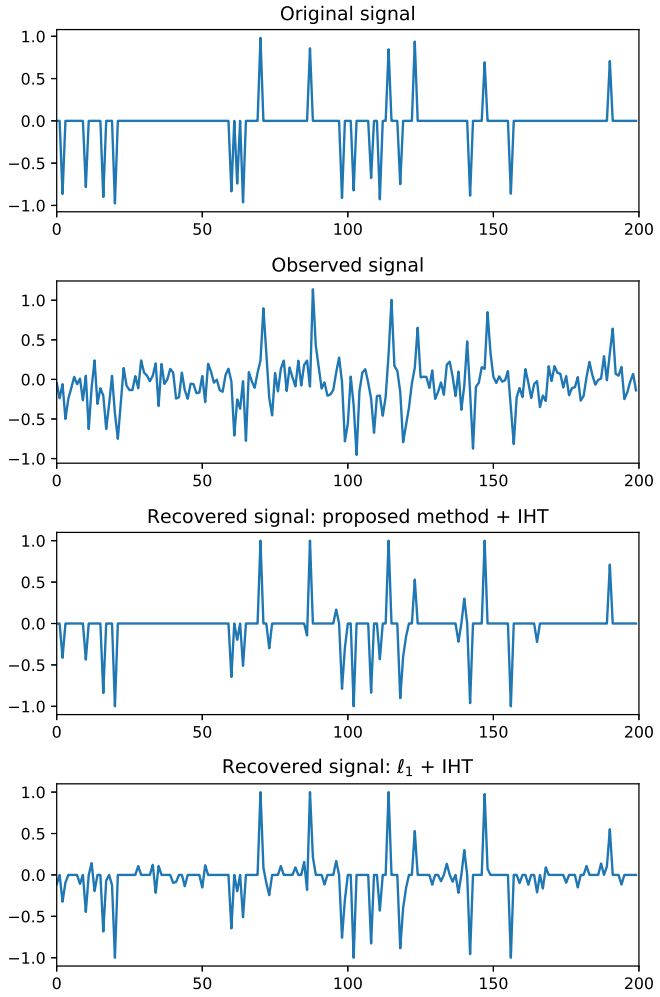

Fig. 6: Typical original signal $\overline{\mathbf{x}}$, observations $\mathbf{d}$ and recovered signal. These results have been obtained with IHT initialized either by our method or by using a linearized model and ℓ_1 penalty.

with ℓ_1 penalty leads to poor results, even when it is associated with an IHT algorithm.

VI. Conclusion

In this paper, we have presented a global optimization approach for addressing a wide range of variational poblems arising in signal processing. More specifically, the proposed method is able to deal with nonlinear models and regularization functions, provided that they can be approximated under a rational form. The validity of the proposed sparse SDP relaxation has been demonstrated on a sparse signal restoration problem where the observations are degraded by a convolution followed by a saturation effect.

This work opens up new perspectives for solving signal recovery and estimation problems where standard optimization algorithms may fail due to the presence of spurious local minimas. On common computer architectures, using existing SDP solvers, the implementation of this approach is however currently limited to relatively small signal dimensions and low filter orders.

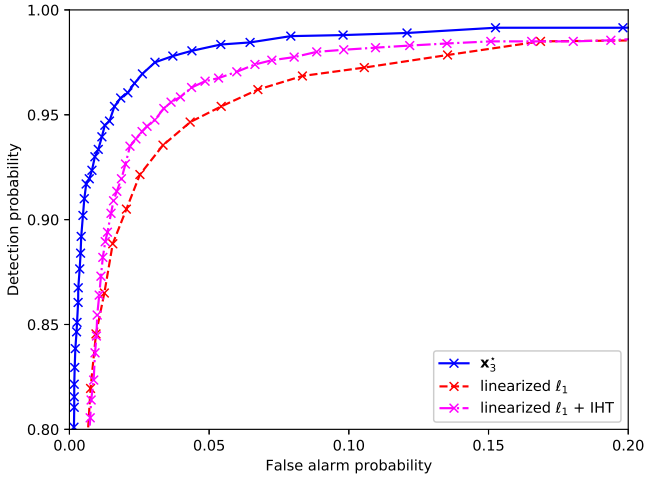

Fig. 7: ROC curve (randomly driven filters, real-valued case, $T = 200$).

References

[1] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] M. Shetzen, *The Volterra and Wiener Theories of Nonlinear Systems*. New York: Wiley and sons, 1980.

[3] N. Dobigeon, J.-Y. Tourneret, C. Richard, J. C. M. Bermudez, S. McLaughlin, and A. O. Hero, "Nonlinear unmixing of hyperspectral images: models and algorithms," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 82–94, Jan. 2014.

[4] Z. Yang, Z. Wang, H. Liu, Y. C. Eldar, and T. Zhang, "Sparse nonlinear regression: Parameter estimation and asymptotic inference," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1511.04514v1

[5] P. T. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. of the 42nd Annual Converence on Information Sciences and Systems*, 2008, pp. 16–21.

[6] Y. Deville and L. T. Duarte, "An overview of blind source separation methods for linear-quadratic and post-nonlinear mixtures," in *Int. Conf. Latent Variable Analysis and Signal Separation*, ser. LNCS, vol. 9237. Liberec, Czech Republic: Springer, 2015, pp. 155–167.

[7] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[8] M. Genzel, "High-dimensional estimation of structured signals from non-linear observations with general convex loss functions," *IEEE Trans. Inform. Theory*, vol. 63, no. 3, pp. 1601–1619, 2017.

[9] M. Genzel and P. Jung, "Sparse recovery from superimposed non-linear sensor measurements," in *Proc. of Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop*, Lisbon, Portugal, June 2017.

[10] Y. Plan and R. Vershynin, "The generalized lasso with non-linear observations," *IEEE Trans. Inform. Theory*, vol. 62, no. 3, pp. 1528–1537, 2016, arXiv: 1502.04071.

[11] M. Nikolova, "Description of the minimizers of least squares regularized with ℓ_0 norm. uniqueness of the global minimizer," *SIAM J. Imaging Sci.*, vol. 6, no. 2, pp. 904–937, 2013.

[12] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 629–654, 2008.

[13] A. Patrascu and I. Necoara, "Random coordinate descent methods for ℓ_0 regularized convex optimization," *IEEE Trans. Automat. Contr.*, vol. 60, no. 7, pp. 1811–1824, Jul. 2015.

[14] S. Geman and D. McClure, "Bayesian image analysis: An application to single photon emission tomography," in *Proc. Statist. Comput. Section Amer. Statist. Association*, 1985, pp. 12–18.

[15] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization," *SIAM J. Imaging Sci.*, vol. 6, no. 1, pp. 563–591, 2013.

[16] A. Florescu, E. Chouzenoux, J.-C. Pesquet, P. Ciuciu, and S. Ciochina, "A majorize-minimize memory gradient method for complex-valued inverse problem," *Signal Process.*, vol. 103, pp. 285–295, Oct. 2014, special issue on Image Restoration and Enhancement: Recent Advances and Applications.

[17] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem," *SIAM J. Imaging Sci.*, vol. 8, no. 3, pp. 1607–1639, 2015.

[18] M. Castella and J.-C. Pesquet, "Optimization of a Geman-McClure like criterion for sparse signal deconvolution," in *Proc. IEEE Int. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Cancun, Mexico, Dec. 2015, pp. 317–320.

[19] ——, "A global optimization approach for rational sparsity promoting criteria," in *Proc. Eur. Signal Image Process. Conf.*, Kos island, Greece, Aug.-Sep. 2017, pp. 166–170.

[20] ——, "Recovery of nonlinearly degraded sparse signals through rational optimization," in *Proc. of Signal Processing with Adaptive Sparse Structured Representations (SPARS) workshop*, Lisbon, Portugal, Jun. 2017.

[21] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, H. H. Bauschke, R. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, Eds. New York: Springer-Verlag, 2010, pp. 185–212.

[22] N. Komodakis and J.-C. Pesquet, "Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems," *IEEE Signal Process. Mag.*, vol. 32, pp. 31–54, Nov. 2015.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Machine Learn.*, vol. 8, no. 1, pp. 1–122, 2011.

[24] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize-minimize subspace strategy for subspace optimization applied to image restoration," *IEEE Trans. Image Process.*, vol. 20, no. 18, pp. 1517–1528, Jun. 2011.

[25] J.-B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, no. 3, pp. 796–817, 2001.

[26] M. Laurent, "Sum of squares, moment matrices and optimization over polynomials," in *Emerging Applications of Algebraic Geometry*, ser. IMA Volumes in Mathematics and its Applications, M. Putinar and S. Sullivant, Eds. Springer, 2009, vol. 149, pp. 157–270.

[27] J.-B. Lasserre, *Moments, Positive Polynomials and Their Applications*, ser. Optimization Series. Imperial College Press, 2010, vol. 1.

[28] D. Jibetean and E. de Klerk, "Global optimization of rational functions: a semidefinite approach," *Math. Progr. (Ser. A)*, no. 106, 2006.

[29] P. A. Parrilo, "Semidefinite programming relaxations for semialgebraic problems," *Math. Programm.*, vol. 96, no. 2, pp. 293–320, 2003. [Online]. Available: https://doi.org/10.1007/s10107-003-0387-5

[30] Z. Luo, W. Ma, A. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, may 2010.

[31] F. Bugarin, D. Henrion, and J.-B. Lasserre, "Minimizing the sum of many rational functions," *Mathematical Programming Computations*, vol. 8, no. 1, pp. 83–111, 2015.

[32] C. Wang, R. Chan, M. Nikolova, R. Plemmons, and S. Prasad, "Nonconvex optimization for 3d point source localization using a rotating point spread function," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1804.04000v1

[33] R. H. Tütüncü, K. C. Toh, and M. J. Tod, "Solving semidefinite-quadratic-linear programs using SDPT3," *Mathematical Programming*, vol. 95, no. 2, pp. 189–217, feb 2003.

[34] M. F. Anjos and J. B. Lasserre, Eds., *Handbook on Semidefinite, Conic and Polynomial Optimization*. Springer US, 2012.

[35] D. Henrion and J.-B. Lasserre, *Positive polynomials in control*, ser. Lecture Notes on Control and Information Sciences. Springer Verlag, Berlin, 2005, vol. 312, ch. Detecting Global Optimality and Extracting Solutions in GloptiPoly.

[36] A. Marmin, M. Castella, J.-C. Pesquet, and L. Duval, "Signal reconstruction from sub-sampled and nonlinearly distorted observations," in *Proc. Eur. Signal Image Process. Conf.*, Rome, Italy, Sep. 2018.

[37] D. Henrion, J.-B. Lasserre, and J. Löfberg, "Gloptipoly3: moments, optimization and semidefinite programming," *Optimization methods and software*, vol. 24, no. 4-5, pp. 761–779, 2009.

[38] C. Chaux, P. L. Combettes, J.-C. Pesquet, and V. R. Wajs, "A variational formulation for frame-based inverse problems," *Inverse Problems*, vol. 23, no. 4, pp. 1495–1518, 2007. [Online]. Available: https://doi.org/10.1088/0266-5611/23/4/008

[39] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods," *Math. Prog.*, vol. 137, no. 1, pp. 91–129, Feb. 2013.